

# MolLoc: a web tool for the local structural alignment of molecular surfaces

Stefano Angaran<sup>1</sup>, Mary Ellen Bock<sup>2</sup>, Claudio Garutti<sup>1,\*</sup> and Concettina Guerra<sup>1,3</sup>

<sup>1</sup>Department of Information Engineering, University of Padova, Via Gradenigo 6a, Padova, Italy,

<sup>2</sup>Department of Statistics, Purdue University 250 N. University Street, West Lafayette, IN 47907-2066 and

<sup>3</sup>College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0280, USA

Received February 28, 2009; Revised April 20, 2009; Accepted May 2, 2009

## ABSTRACT

**MolLoc stands for Molecular Local surface comparison, and is a web server for the structural comparison of molecular surfaces. Given two structures in PDB format, the user can compare their binding sites, cavities or any arbitrary residue selection. Moreover, the web server allows the comparison of a query structure with a list of structures. Each comparison produces a structural alignment that maximizes the extension of the superimposition of the surfaces, and returns the pairs of atoms with similar physicochemical properties that are close in space after the superimposition. Based on this subset of atoms sharing similar physicochemical properties a new rototranslation is derived that best superimposes them. MolLoc approach is both local and surface-oriented, and therefore it can be particularly useful when testing if molecules with different sequences and folds share any local surface similarity. The MolLoc web server is available at <http://bcb.dei.unipd.it/MolLoc>.**

## INTRODUCTION

Structural comparison is used extensively to determine the function of proteins, and to study the interactions between proteins and nucleic acids. Most of the tools currently available are for global structural comparison. For instance, SSAP (1), STRUCTAL (2), DALI (3), LSQMAN (4), CE (5) and SSM (6) find the rototranslation of two given structures that maximizes the number of residues that are close after the global alignment (7). Moreover, the molecule is often represented by its atoms or a subset of its atoms (e.g.  $C_\alpha$  atoms), which is a simplified representation of the molecular structure.

In this article, we introduce MolLoc (Molecular Local), a web server for the recognition of similar regions on molecular surfaces. The surfaces may be restricted to cavities, binding sites or any residue selection of complete

protein, RNA or DNA. The server determines the most extended similar regions of the selected surfaces. This application can be particularly useful when the user is interested in inferring functional information for a molecule, be it a protein, RNA or DNA. First, if the molecule has a binding site, the surface comparison of its binding site with binding sites of other molecules can identify potential ligands or inhibitors to use within its binding site. Second, if the molecule has no known binding sites but has a set of functionally relevant residues, the comparison of these residues with other binding sites can suggest new ligands for these residues. Third, if the molecule has no functional characterization at all, comparing its cavities with other binding sites can provide clues to the molecular function, since binding sites usually lie in cavities (8,9).

Available tools that provide related facilities are Multibind (10), 3D-surfer (11), eF-seek (12) and FunClust (13). Multibind (<http://bioinfo3d.cs.tau.ac.il/MultiBind/>) recognizes spatial chemical binding patterns common to a set of protein structures. It handles several proteins at once but, like eF-seek (<http://ef-site.hgc.jp/eF-seek/index.jsp>), aligns binding sites only. 3D-surfer (<http://dragon.bio.purdue.edu/3d-surfer/>) performs a comparison of a query protein surface against all protein structures in the PDB and retrieves those with highest global surface similarity with the query. It establishes global surfaces similarity but, unlike MolLoc, does not search for local surface regions corresponding to candidate binding sites. FunClust (<http://pdbsfun.uniroma2.it/funclust/>) takes as input a list of proteins and identifies a set of shared residues. It matches proteins based on a local structural representation, not on surface information as in MolLoc.

## MATERIALS AND METHODS

### Input data

MolLoc can perform pairwise surface comparison of two structures, or multiple pairwise surface comparison of a query structure with a list of structures (Figure 1).

\*To whom correspondence should be addressed. Tel: +39 049 827 7928; Fax: +39 049 827 7799; Email: garuttic@dei.unipd.it

[ Home - Publications - Help - Contact ]

**Pairwise Surface Comparison (one vs one)**

PDB code 1 ⓘ (e.g. 1atp)  or upload file in PDB format ⓘ

PDB code 2 ⓘ (e.g. 1csn)  or upload file in PDB format ⓘ

Email ⓘ results to  (optional)

**Multiple Surface Comparison (one vs many)**

Query PDB code ⓘ (e.g. 1atp)  or upload file in PDB format ⓘ

List of PDB codes (max 20) ⓘ  
(e.g. 1csn 1phk,A 1mjh,AB)

Email ⓘ results to  (mandatory)

**Figure 1.** Home page of MolLoc. The user can run a pairwise surface comparison between two structures, or a multiple pairwise surface comparison between a query structure and a list of structures from the PDB. In the latter case, the email address is mandatory.

Select Atoms for 1atp,E

Restrict to Binding Site(s) ⓘ

- All
- MN:\_\_:2
- MN:\_\_:3
- PO3:E:197
- PO3:E:338
- ATP:\_\_:1

Restrict to Cavities ⓘ

- shallow cavities
- recommended
- deep cavities
- very deep cavities

Restrict to Residue Numbers  Set  
( e.g. 24, 37-39, 42 )

Selection Preview

Jmol

**Figure 2.** Atoms selection. The user has to provide a nonempty selection of atoms for each structure.

**Pairwise surface comparison.** In the pairwise surface comparison, MolLoc takes in input the coordinate files of two molecules in PDB format (14). The user can either insert the pdb-ids of structures that belong to the PDB or upload his/her own structures. Optionally, the user can write his/her email address, to receive a link to the results at the end of the computation. Next, the user specifies one or more chains from each structure. The third step is the selection of the regions to compare (Figure 2). For each structure, the user can select one or more binding sites, cavities or any set of residues. A ligand is an HETATM residue different than HOH in the pdb file, and the binding site for a ligand  $L$  in a structure  $S$  is defined as the subset of atoms of  $S$  that are closer than  $6\text{ \AA}$  to at

least one atom of  $L$ . The cavities are generated with different depths, ranging from shallow cavities to very deep cavities. Each structure is associated to a Jmol visualization (15), thus enabling the user to visualize the selected regions. Finally, the user can choose the comparison method. The method called *only geometry* does not make use of any physicochemical property, while the method called *geometry + atomtype* starts from the roto-translation obtained with the purely geometrical method and iteratively refines it by matching pairs of atoms with the same atomtype, as specified in (16). The atomtypes are defined only for protein atoms, and therefore the second method is specific for the comparison of two proteins.

**Multiple pairwise surface comparison.** MolLoc also allows multiple pairwise surface comparison of one query structure with up to 20 other structures. Here, the email address is mandatory. For the query structure, the user can still specify a PDB ID or upload a structure. The preprocessing phase (chain selection and atoms selection) works like in the pairwise case. The list of other structures must belong to the PDB, and the user can specify the chain(s) using the syntax `pdb_id, chain(s)` (e.g. 1atp, E1). For these other structures, MolLoc automatically selects all of their binding sites for comparison with the query structure.

### Processing method

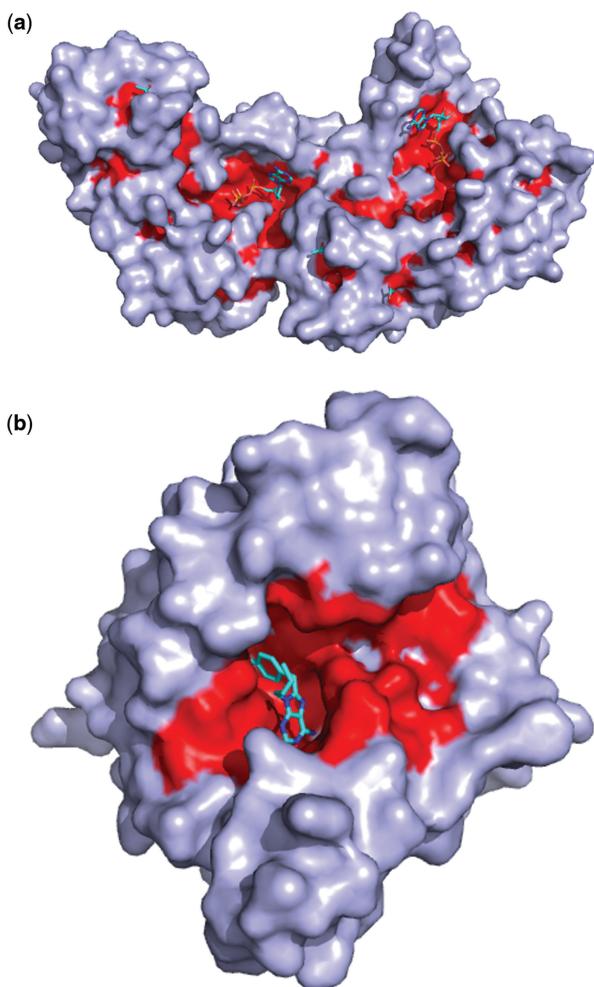
The web server is built upon a method for the discovery of similar regions on two molecular surfaces based on a spin-image representation of the surfaces (17). Given two structures, the *only geometry* method:

- (1) builds their Connolly's molecular surfaces (18);
- (2) builds the spin-image representations of Connolly's points (19);
- (3) compares the spin images of the atoms of the two surfaces, and puts them in correspondence if their correlation is high ( $>0.5$ );
- (4) finds sets of geometrically consistent correspondences using a greedy procedure;
- (5) the largest set of correspondences represents the best solution;
- (6) the obtained point correspondences are given as input to the Horn method (20) that produces the best rototranslation that superimposes the two regions.

The *geometry + atomtype* method takes as input the superimposition obtained with the *only geometry* method, and checks for atoms of the first structure that are closer than  $2.5\text{\AA}$  to at least one atom with the same atomtype (16) belonging to the second structure. The atomtypes are defined for protein atoms only, and correspond to the following properties: hydrogen-bond donor (DON), hydrogen-bond acceptor (ACC), mixed donor/acceptor (DAC), hydrophobic aliphatic (ALI) and aromatic contacts (PI). Then, the method keeps all the pairs of atoms that are unambiguous, where a pair  $(A_i, B_j)$  of atoms  $A_i$  in the first structure and  $B_j$  in the second structure is unambiguous if  $B_j$  is the only atom that is closer than  $2.5\text{\AA}$  to  $A_i$ , and *vice versa*. These  $n$  pairs are given in input to the Horn's method, that produces a second rototranslation. Again, the method checks for unambiguous atom pairs. If the number  $m$  of the new set of pairs is such that  $m > n$ , then the procedure iterates (for a maximum of 10 steps), else it stops.

The cavity detection procedure is a novel method that allows the fast determination of cavities with adjustable depth. The method consists of the following steps:

- (1) for each atom  $i$  belonging to the surface, count the number of atom centers  $N_C(i)$  that lie within a radius  $R$  from the center of  $i$ ;



**Figure 3.** (a) Cavities of cystic fibrosis transmembrane conductance regulator with  $R = 12$ ; (b) cavity of the human chaperon hsp90 with  $R = 20$ .

- (2) the cavity atoms are defined as those atoms  $i$  s.t.  $N_C(i) \geq \mu(N_C) + \xi\sigma(N_C)$ , where  $\mu(N_C)$  is the mean,  $\sigma(N_C)$  is the standard deviation and  $\xi = 0.5$ .

The cavity size and topology depends on the value of radius  $R$  (Figure 3). In MolLoc, the choices are  $R = 4, 8, 12, 20$ , where  $R = 4$  is for shallow cavities and  $R = 20$  is for very deep cavities.

### Output of the web server

The top of the results page presents the statistics of the experiment: number of selected input atoms in each structure, ratio of corresponding surface area to the selected input surface area in each structure, number of corresponding atoms with the same atomtype between the two structures, together with their RMSD. Furthermore, the page allows the download of the first PDB file by clicking on its protein name at the top of the right-hand column, the second PDB file rototranslated after the superimposition by clicking on its protein name in the right-hand column, the matrix of rototranslation in

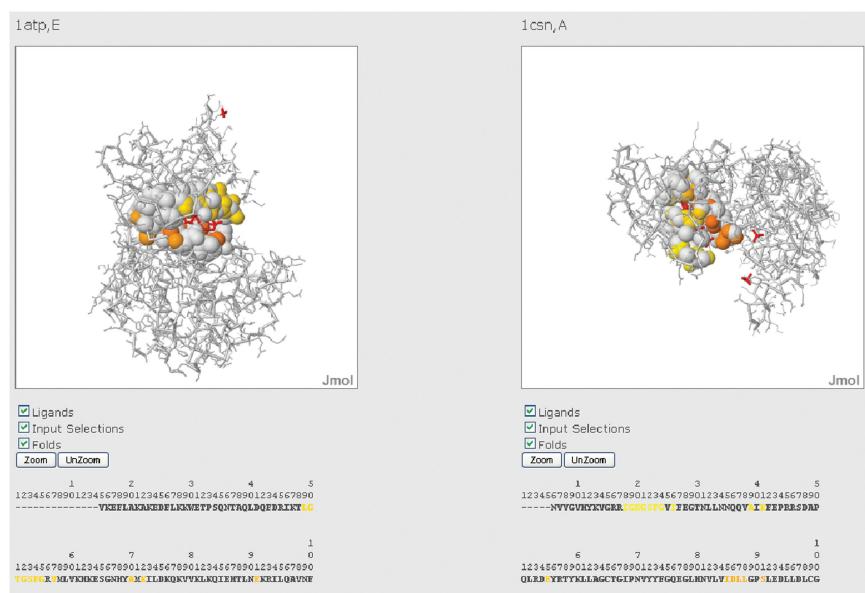
DaliLite (21) format, the table of atom correspondences and a PyMol (22) script that shows the pairs of surface points that generated the superimposition of the second structure onto the first.

Below, the results page (Figure 4) presents a Jmol visualization of the two superimposed structures, and a table containing the correspondences between the atoms of the

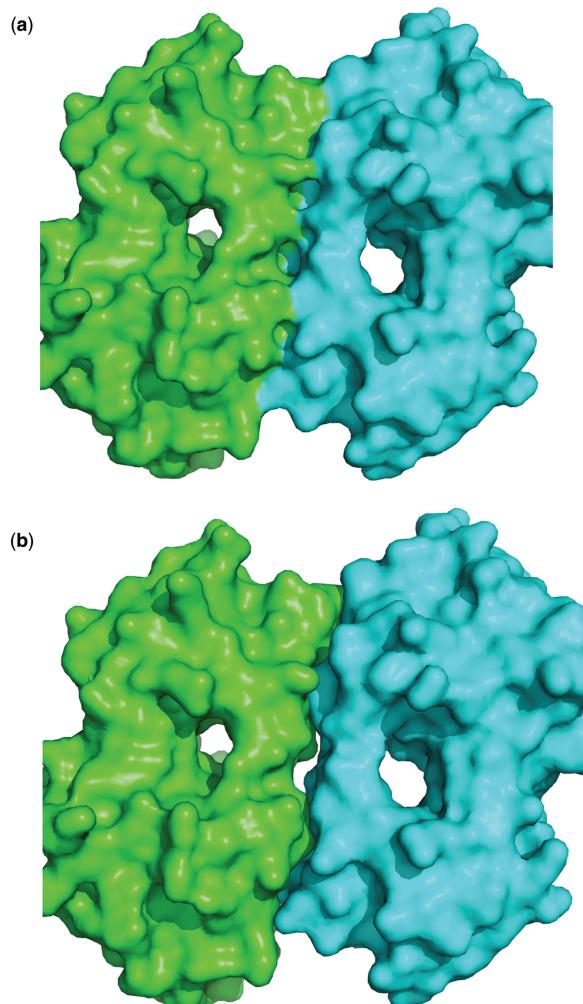
two structures with the same atomtype and closer than 2.5 Å after the superimposition. Each pair of corresponding atoms can be visualized in spacefill by clicking on its check box in the right-hand column of the table. (All the corresponding atoms can be simultaneously selected by clicking at the top of this column.) There is also a side-by-side view of the two structures where the regions that

1atp,E				1csn,A				Check to show atoms in the picture	
ATOM ▲▼	RES ▲▼	CHAIN ▲▼	RES NUM ▲▼	ATOM ▲▼	RES ▲▼	CHAIN ▲▼	RES NUM ▲▼	TYPE ▲▼	<input type="checkbox"/>
O	LEU	E	49	O	ILE	A	18	AC	<input type="checkbox"/>
C	LEU	E	49	C	ILE	A	18	PI	<input type="checkbox"/>
C	GLY	E	50	C	GLY	A	19	PI	<input type="checkbox"/>
C	THR	E	51	C	GLU	A	20	PI	<input type="checkbox"/>
O	THR	E	51	O	GLU	A	20	AC	<input type="checkbox"/>
N	THR	E	51	N	GLU	A	20	DO	<input type="checkbox"/>
O	GLY	E	52	O	GLY	A	21	AC	<input type="checkbox"/>
C	GLY	E	52	C	GLY	A	21	PI	<input type="checkbox"/>
OG	SER	E	53	OG	SER	A	22	DA	<input type="checkbox"/>
O	SER	E	53	O	SER	A	22	AC	<input type="checkbox"/>
N	SER	E	53	N	SER	A	22	DO	<input type="checkbox"/>
N	PHE	E	54	N	PHE	A	23	DO	<input type="checkbox"/>
C	PHE	E	54	C	PHE	A	23	PI	<input type="checkbox"/>
O	PHE	E	54	O	PHE	A	23	AC	<input type="checkbox"/>
N	GLY	E	55	N	GLY	A	24	DO	<input type="checkbox"/>

**Figure 4.** Alignment obtained from the comparison between the binding sites of 1atp, chain E and of 1csn, chain A. The table on the left shows the atom correspondences. When a box is checked, the relative pair of atoms is shown in spacefill.



**Figure 5.** Side-by-side view of 1atp, chain E and 1csn, chain A. The atoms that belong to the solution are colored following a gradient from the N-terminal (yellow) to the C-terminal (red). For each structure, the residues that belong to the solution are colored according to the colors in the Imol visualization.



**Figure 6.** PDB structure 104l. (a) Surface representation of the two chains together. (b) Surface representation of the two chains separately. The two representations differ in the interface between the two chains.

are found as similar by the comparison are colored with a gradient from Nter to Cter (Figure 5). Below the two structures are the residues of the two sequences and the residues of each sequence that belong to the solution have the same color as in the Jmol visualization.

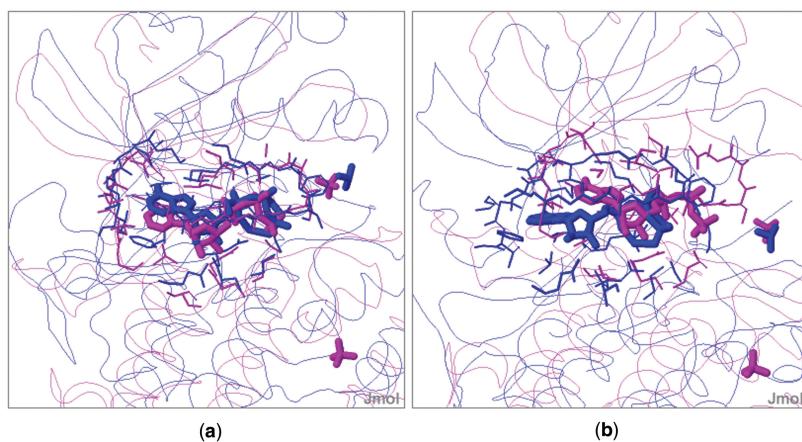
The results of the multiple pairwise surface comparison are summarized in a table sorted by number of corresponding atoms for each pair. Each structure in the left-hand column is linked with the page that stores the results of the comparison between that structure and the query structure.

#### A note on chain selection

Users have to keep in mind a caveat when dealing with chain selection for a structure with multiple chains. That is, the surface representation of two contiguous chains is different from the surface representation of each of the two chains separately (Figure 6), and therefore the result of the comparison can be different. For example, the molecular surface of 1atp, chain E, is different from the molecular surface of 1atp, chains E and I. In fact, the ATP binding pocket of 1atp,E is an open cavity, while the ATP binding pocket of 1atp, EI is an internal cavity. Hence, the two surfaces are different and the optimal alignment between the ATP binding pockets of 1atp, EI (both chains) and of 1csn,A is slightly different from the optimal alignment between the ATP binding pockets of 1atp,E (only one chain) and of 1csn,A (Figure 7). In this example, the solution of the comparison between 1atp, EI and 1csn,A contains only corresponding atoms from chain E on structure 1atp. The web server gives a warning message, telling the user that the alignment may change if run again on a single chain (in this case chain E) of a multiple chain protein.

#### PERFORMANCE

MolLoc web server uses several different software modules to build the surface representation, find the binding sites and the cavities, and to compare the surfaces.



**Figure 7.** 1atp in purple, 1csn in blue. (a) The result of the comparison between 1atp,E and 1csn,A. (b) The result of the comparison of both chains of 1atp (E and I) with 1csn,A. In this case, differences in the initial surface lead to differences in the surface alignment.

The surface determination routine scales linearly with the number of atoms, ranging from a few seconds for ordinary structures (e.g. 1atp, chain E) to minutes for huge macromolecular complexes (e.g. 1aei, all chains). Therefore, each time the user provides one or more new chains from the PDB, their surface representations are saved into an internal database, to avoid rebuilding when the same chains are invoked again.

The surface comparison routine is the most time-consuming module on the web server. Its time complexity is  $O(n \times m)$ , where  $n, m$  are the number of atoms selected in the first and the second structure. The execution time ranges from a few seconds for comparison of medium-sized binding sites (e.g. the binding site of ATP in 1atp,E with the binding site of ATP in 1csn,A) to minutes for comparison of extended areas (e.g. residues 15–350 of 1atp,E with residues 6–298 of 1csn,A).

## CONCLUSION

We have presented MolLoc, a new server for the structural comparison of molecular surfaces. The server allows comparison of binding sites, cavities and any arbitrary residue selection. The adopted approach is both local and surface-oriented, and therefore it can be particularly useful when testing if molecules with different sequences and folds share any local surface similarity.

## FUNDING

Funding for open access charge: Progetto di Ateneo, Universita' degli Studi di Padova, and Progetto CARIPARO, Padova.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.
2. Gerstein,M. and Levitt,M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.*, **7**, 445–456.
3. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
4. Kleywegt,G.J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D*, **D52**, 842–857.
5. Shindyalov,I.N. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng. Des. Sel.*, **11**, 739–747.
6. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography* Vol. 60, International Union of Crystallography, pp. 2256–2268.
7. Novotny,M., Madsen,D. and Kleywegt,G.J. (2004) Evaluation of protein fold comparison servers. *Proteins Struct. Funct. Bioinform.*, **54**, 260–270.
8. Glaser,F., Morris,R.J., Najmanovich,R.J., Laskowski,R.A. and Thornton,J.M. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins. Struct. Funct. Bioinform.*, **62**, 479–488.
9. Bock,M.E., Garutti,C. and Guerra,C. (2007) Effective labeling of molecular surface points for cavity detection and location of putative binding sites. *Computational Systems Bioinformatics: Proceedings of the CSB 2007 Conference*, Imperial College Press, London, pp. 263–274.
10. Shatsky,M., Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2005) Recognition of binding patterns common to a set of protein structures. *Lecture Notes in Computer Science*. Vol. 3500, Springer, pp. 440–455.
11. Sael,L., La,D., Li,B., Rustamov,R. and Kihara,D. (2008) Rapid comparison of properties on protein surface. *Proteins Struct. Funct. Bioinform.*, **73**, 1–10.
12. Kinoshita,K., Murakami,Y. and Nakamura,H. (2007) eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.*, **35**, W398–W402.
13. Ausiello,G., Gherardini,P.F., Marcatili,P., Tramontano,A., Via,A. and Helmer-Citterich,M. (2008) FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics*, **9**, S2.
14. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. et al. (2002) The protein data bank. *Acta Crystallogr. D*, **D58**, 899–907.
15. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
16. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
17. Bock,M.E., Garutti,C. and Guerra,C. (2007) Discovery of similar regions on protein surfaces. *J. Comput. Biol.*, **14**, 285–299.
18. Connolly,M.L. (1983) Analytical molecular surface calculation. *J. Appl. Crystallogr.*, **16**, 548–558.
19. Johnson,A.E. and Hebert,M. (1999) Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, **21**, 433–449.
20. Horn,B.K.P. (1987) Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A*, **4**, 629–642.
21. Holm,L. and Park,J. (2000) DALI Lite workbench for protein structure comparison. *Bioinformatics*. Vol. 16, Oxford University Press, pp. 566–567.
22. DeLano,W.L. (2002) The PyMOL molecular graphics system, DeLano Scientific, San Carlos, CA, USA, Available at <http://pymol.sourceforge.net/faq.html#CITE>.